An Information-Theoretic Approach to Modeling Diachronic Change in Scientific English

Stefania Degaetano-Ortlieb, Hannah Kermes, Ashraf Khamis, Elke Teich

Universität des Saarlandes

Universität Campus A2.2, 66123 Saarbrücken, Germany

{*s.degaetano, h.kermes, e.teich*}*@mx.uni-saarland.de, ashraf.khamis@uni-saarland.de*

Abstract

We present an information-theoretic approach to investigate diachronic change in scientific English. Our main assumption is that over time scientific English has become increasingly dense, i.e. linguistic constructions allowing dense packing of information are progressively used.

So far, diachronic change in scientific writing has been investigated by means of frequency-based approaches (see e.g. Halliday (1988); Atkinson (1998); Biber (2006b,c); Biber and Gray (2016); Banks (2008); Taavitsainen and Pahta (2010)). We use information-theoretic measures (entropy, surprisal; (Shannon, 1949)) to assess features previously stated to change over time and to discover new, latent features from the data itself that are involved in diachronic change.

For this, we use the Royal Society Corpus (RSC) (Kermes et al., 2016), which spans over the time period 1665 to 1869. We present three kinds of analyses: nominal compounding (typical of academic writing), modal verbs (shown to have changed in frequency over time), and an analysis based on part-of-speech trigrams to detect new features that change diachronically. We show how information-theoretic measures help to investigate, evaluate and detect features involved in diachronic change.

Keywords

Scientific English, diachronic change, information theory, entropy, surprisal

1 Introduction

We report on a project investigating the diachronic development of English scientific writing from the mid-17th to the mid-19th century. While scientific discourse is a much researched topic in corpus linguistics (e.g. Biber (2006b,a); Biber and Gray (2016, 2013)), most works focus on earlier periods (notably Early Modern English; e.g. Taavitsainen et al. (2011); Taavitsainen and Pahta (2010)) or contemporary writing (e.g. Biber (2006c)).

In the period we are dealing with, which is marked by the transition from Early Modern to Late Modern English, scientific activity became increasingly diversified and specialized as well as professionalized and institutionalized (the major modern scientific disciplines developed during that time). We are interested in the linguistic effects of the processes of diversification/specialization and professionalization/institutionalization. More concretely, we assume that for scientific communication to become fully functional for experts, it needed to develop an efficient code, i.e. a code with minimal redundancy and sufficient expressivity. Specifically, we pursue the following hypotheses:

- as an effect of diversification/specialization, scientific texts will exhibit a *greater encoding density* over time (Halliday and Martin, 1993, 54–68), i.e. linguistic constructions allowing denser information packing will be increasingly used;
- as an effect of professionalization/institutionalization, scientific texts will become more standardized over time, exhibiting *greater linguistic uniformity*, i.e. the linguistic forms used will be increasingly conventionalized.

We further assume that these effects are measurable in the linguistic signal using information-theoretic methods (cf. Shannon (1949)), in particular measures based on entropy and surprisal (cf. Crocker et al. (2015)). Here, the main challenges are (a) to identify those linguistic features that contribute to making scientific writing a distinctive type of discourse and (b) to assess the distinctive force of those features. A variety of features have been looked at in previous work on scientific writing, including high lexical and term densities, low type-token ratio, nominal groups with extensive pre- and postmodification or reduced relative clauses (Halliday, 1988; Biber, 2006b,c). However, we suspect that there are quite a few latent patterns that have yet to be uncovered. Also, most corpus-linguistic approaches are based on frequency (unconditioned probabilities), diachronic change being reported as change in frequency distributions of single items or constructions. Context is not taken into account systematically (except for collocations and lexical bundles). Also, frequencies alone cannot be directly interpreted in terms of typicality (or changes in what is typical in a given time period vs. another).

Information-theoretic measures, instead, are based on conditional probabilities, diachronic changes can be represented by changes of probabilities *in context*. Consider, for example, the decrease of modal verbs in English in general (cf. Leech (2003)). As Leech mentions himself, beside obtaining evidence of frequency-based change, one is ultimately interested in finding out how and why the change has been taking place (cf. Leech (2003, 232)): Does the change affect all modal verbs or only particular ones? And is the change context-independent or is it primed by particular contexts? While in a frequency-based approach, context may be explored in a separate, secondary step, information-theoretic measures such as entropy and surprisal inherently model context.

We present a two-pronged approach to the detection and analysis of features involved in diachronic change in scientific English. Using entropy-based models, we newly assess features that have previously been shown to be involved in diachronic change in scientific English in frequency-based accounts, on the one hand, and we detect new, latent features and evaluate them using the same kind of models, on the other hand. In terms of linguistic theory, we are committed to Hallidayan register theory (Halliday and Hasan, 1985) which states that linguistic variation is driven by settings in situational context in terms of field, tenor and mode of discourse.

We proceed as follows. First, we introduce the data we use for our investigation, the Royal Society Corpus (RSC) (Kermes et al., 2016) (Section 2). Second, we introduce our methods, notably entropy and surprisal-based language modeling (Section 3). In Section 4, we present three kinds of analyses using the RSC: (A1) nominal compounding — a feature that has been reported as typical of scientific text and is related to the field of discourse, (A2) modal verbs — a feature that has been shown before to be involved in diachronic change and which is related to the tenor of discourse, and (A3) an exploratory analysis based on trigrams for detection of new features. Section 5 concludes with a summary and discussion.

journal	**period**	**text type**				
		book reviews	**articles**	**miscellaneous**	**obituaries**	**total**
Philosophical Transactions	1665–1678	124	641	154	–	919
Philosophical Transactions	1683–1775	154	3,903	338	–	4,395
Philosophical Transactions of the Royal Society of London (PTRSL)	1776–1869	–	2,531	283	–	2,814
Abstracts of Papers Printed in PTRSL	1800–1842	–	1,316	15	–	1,331
Abstracts of Papers Communicated to RSL	1843–1861	–	429	5	–	434
Proceedings of RSL	1862–1869	–	1,476	38	14	1,528
Total		278	10,296	833	14	11,421

Table 1: Material used for the RSC

2 Data

The Royal Society Corpus (RSC) contains texts from the first two centuries of publications by the Royal Society of London (1665–1869) (cf. Table 1 for an overview). We obtained the material from JSTOR[1] in a well-formed XML format including meta-data (e.g. author(s), text type (such as article, abstract), day, month and year of publication, volume, text ID, and title).

The corpus-building process is inspired by the idea of Agile Software Development (Cockburn, 2001) according to which new, improved versions of a piece of software are produced continuously. In our case, we intertwine corpus building, corpus annotation and analysis to produce new versions of the corpus whenever we encounter problems in data quality. Although already digitized, the source texts contained a considerable amount of noise, e.g. OCR errors and foreign language material (Latin, French, Italian a.o.), which can impact the quality of any step in corpus processing as well as corpus analysis. We apply a dedicated pipeline for corpus building divided into three main steps: (i) preprocessing, (ii) linguistic annotation, and (iii) corpus encoding. The steps in the pipeline are mostly automatic; manual work is kept to a minimum and is applied prior to the first automatic step in the pipeline. The scripts we use for processing are adapted to the special requirements of the source data. They include transformation of data into a standardized format, reduction of noise and derivation and annotation of meta-data. The main types of noise reduction that we address are OCR errors, layout problems and foreign language material. Sources for relevant meta-data are: (i) the given meta-data, (ii) (lexical) triggers in the texts, (iii) a combination of (i) and (ii), (iv) results of pattern-based and/or data-mining techniques.

For the time being, linguistic annotation is mainly performed on the token level. We annotate words (normalized and original word forms), lemmas and parts of speech. For the linguistic annotation we use existing tools: VARD (Baron and Rayson, 2008) for normalization and TreeTagger (Schmid, 1994, 1995) for tokenization, lemmatization and part-of-speech (POS) tagging. For the training and evaluation of VARD, we created a manually annotated (normalization, part-of-speech tags) subcorpus of the RSC ($\sim$56.000 tokens) and divided it into roughly equal-sized subsets. The trained version of VARD exhibited an increase of more than 10% in precision (61.8% to 72.8%) and almost double the recall (31.3% to 57.7%). For the evaluation of TreeTagger, we used the whole subcorpus (precision: 94.5% on original and 95.1% on normalized word forms).

We encode the corpus in CQP format (CWB; Evert and Hardie, 2011) for corpus query and analysis. Currently, we annotate lemmas, normalized/original forms, POS tags as well as entropy values (cf. Section 3 be-

[1] http://www.jstor.org/

low). Additionally, the format allows for structural information in the form of XML tags with attribute-value pairs. After encoding, the corpus may be queried on the command-line or using a web-based GUI (CQPweb (Hardie, 2012)). For diachronic analysis, we have divided the corpus into slices of one year, ten years and approximately fifty years (labeled as follows: *1650*: 1665–1699, *1700*: 1700–1749, *1750*: 1750–1799, *1800*: 1800–1849, *1850*: 1850–1869). See also Kermes et al. (2016) for a more detailed description of the corpus-building process.

3 Analytical Methods

In this section, we present the information-theoretic measures we use to detect and analyze features of diachronic change.

3.1 *Surprisal and Entropy*

Surprisal is a measure of information calculating the number of bits used to encode a message. Applied to language, the number of bits being transmitted by a particular linguistic unit (word, syllable, phrase, etc.) in a running text or stream of speech is dependent on that unit's probability in context — formally $p(unit|context)$. Context can relate here to the context of the preceding unit(s)[2], the wider context of a stretch of text, a whole text or a set of texts[3]. Simply put, the more probable a linguistic unit is in a particular context, with an optimal encoding the fewer bits are used to encode it (or, in other words, the less surprising/informative it will be) and vice versa, the less probable a linguistic unit is in a particular context, the more bits are used to encode it (the more surprising/informative it will be). Formally, surprisal is quantified as the negative log probability of a unit (e.g. a word) in context (e.g. its preceding words):

$$S(unit) = -log_2 p(unit|context)$$

For illustration, consider the following examples:

(1) *John accidentally mailed the letter without a stamp.*

(2) *John went to the shop to buy a stamp.*

Comparing (1) to (2), *stamp* is a much more expected, probable continuation of *John accidentally mailed the letter without a* than of *John went to the shop to buy a.* Assume, for instance, that the only possible continuations for (1) are *stamp* and *ZIP-code* and that they are equally likely. The probability

[2] As used e.g. in part-of-speech tagging (Manning and Schütze, 2001, chap. 10)

[3] As used e.g. in topic modeling (Blei et al., 2003)

for *stamp* would be 1 over 2 (i.e. 0.5). The amount of bits needed to encode *stamp* in (1) would be:

$$S(stamp) = -log_2 p(stamp|John\ accidentally\ mailed\ the\ letter\ without\ a)$$

$$= -log_2 p(0.5) = 1\ bit$$

If there were 10 possible continuations for (2) which are equally likely, the probability of *stamp* would be 1 over 10 (i.e. 0.1). Thus, the amount of bits needed to encode *stamp* in (2) would be:

$$S(stamp) = -log_2 p(stamp|John\ went\ to\ the\ shop\ to\ buy\ a)$$

$$= -log_2 p(0.1) = 3.32\ bits$$

So fewer bits are needed to encode *stamp* in (1) vs. (2) (compare 1 bit vs. 3.32 bits). The more likely case, however, is that the distribution is skewed, some options being more probable than others. If *stamp* in (1) was more likely than *ZIP-Code*, say with a probability of 0.7, then the amount of bits needed to encode stamp in (1) would be:

$$S(stamp) = -log_2 p(stamp|John\ accidentally\ mailed\ the\ letter\ without\ a)$$

$$= -log_2 p(0.7) = 0.51\ bits$$

This is intuitive, as *stamp* has a higher probability to occur and thus the uncertainty of this continuation is lower (i.e. the entropy is lower) than with a balanced distribution (i.e. with equally likely options).

Typically, in the analysis of texts or corpora, we are not interested in the surprisal of just one occurrence of a particular unit but *all* its occurrences, i.e. its *average surprisal*:

$$AvS(unit) = \frac{1}{|unit_i|} * -\sum_i log_2 p(unit_i|context_i)$$

where $|unit_i|$ denotes the number of occurrences of a unit. For our above example of *stamp* in (1) the AvS for *stamp* occurring 7 times in a corpus and twice in the context shown in (1) would be:

$$AvS(stamp) = \frac{1}{|7|} * -(log_2 p(0.5) + log_2 p(0.5)) = 0.95\ bits$$

The notion of average surprisal is immediately relevant for our hypothesis of increasing encoding density in relation to specialization/diversification and is applied in analyses (A1) and (A2) below, focusing on words (unigrams) as units and their preceding word context (Sections 4.1 and 4.2). Also, it has been shown that the more predictable (low in surprisal) a unit is, the shorter its linguistic encoding will be. Cases in point are reduced vs. full relative clauses (Jaeger, 2011), shorter vs. longer word durations (Sayeed et al., 2015) or the marking of discourse relations (Asr and Demberg, 2013). In analysis (A1), we will consider this notion of shorter vs. longer linguistic encoding by comparing noun-noun compounds with their prepositional phrase counterparts (Section 4.1).

Note that when applied to all different units (e.g. the words in a text or corpus), average surprisal is equivalent to entropy (cf. Genzel and Charniak's entropy rate (Genzel and Charniak, 2002)):

$$H = -\sum_{i} p(unit_i|context_i) log_2 p(unit_i|context_i)$$

Overall, the concepts of surprisal and entropy fit very well with that of language use as choice in context (cf. Crocker et al. (2015)), as formulated in many functionalist approaches to language, be that context the cotext (as e.g. for collocations (Firth, 1957)) or the context of situation (Halliday, 1985). Choice in context can thus be appropriately modeled on this basis.

3.2 *Cross-entropy and Relative Entropy*

For our comparative analysis (A3), we need a slightly different perspective. First, we want to look at trigrams rather than words; second, we are interested in their *relative* contribution to diachronic distinction. This means that we need a method to compare probability distributions across time periods to see whether they are different or not. Thus, also the notion of context differs here: rather than referring to the preceding words of a unit, it refers to a time period.

One measure to this end that is often applied in computational language modeling is cross-entropy:

$$H(context1; context2) = -\sum_{i} p(unit_i|context1) log_2 p(unit_i|context2)$$

which gives the *average number of bits* needed to encode a unit when a non-optimal model is used for encoding. Cross-entropy is commonly used for the quality assessment of language models, comparing a model trained on one set of data (training data) on another set of data (test data). The smaller the difference (in bits), the better the model is said to be.

For our purposes we need a slightly different concept, namely that of *relative entropy* which refers to the number of *additional bits* needed when a non-optimal encoding is used. This is formalized by Kullback-Leibler Divergence (KLD), which captures the difference (in number of bits) between the cross-entropy between two data sets A and B and the entropy of A alone, i.e. $H(A; B) - H(A)$. For two time periods T1 and T2, we can thus use KLD as follows:

$$D_{KL}(T1||T2) = -\sum_{i} p(unit_i|T1) log_2 \frac{p(unit_i|T1)}{p(unit_i|T2)}$$

On this basis, the more additional bits are needed for encoding a given unit, the more distinctive (and thus typical) that unit (feature) is for a given time period vs. another time period (cf. Fankhauser et al. (2014)).

We use KLD in analysis (A3), taking trigrams as units to discover new, latent features involved in diachronic change and different time periods as context (for details see Section 4.3).

4 Analyses

4.1 *Compounds vs. Prepositional Phrases (A1)*

A field-related feature we examine involves the possible alternation between common N-N compounds (e.g. *copper alloy*) and their N-PREP-N counterparts (e.g. *alloy of copper*).[4] To our best knowledge, there has been no clear distinction in the literature between pre- and postmodification patterns in the noun phrase (NP) with regard to encoding density. Comparing structural complexity across registers, Biber (1988) drew a parallel between clausal subordination in speech and NP 'heaviness' (Aarts, 1992: 83) in formal writing as two mechanisms for denser encoding and pointed out that 60% of NPs in academic writing are pre- or postmodified (cf. Biber et al., 1999: 578).

We analyze 139 N-N compounds and their exact N-PREP-N counterparts in the RSC to investigate diachronic changes in their frequency distribution, assuming an increase in compound use and a decrease in the prepositional counterpart (cf. Leech and Smith (2009) and Hundt et al. (2012)), while taking into account their average surprisal and syntagmatic context.

For this, we extract any two-noun lemma sequences (excluding preceding or following nouns) forming compounds whose heads occur at least 10 times in the RSC. We then query the corpus for their N-PREP-(DET)-N counterparts (again, excluding any preceding or following nouns) so that we have the exact head lemmas in both N-N compounds and N-PREP-N constructions. That has produced a total of 53 matching heads and 139 variation patterns in the period 1665–1869.

Diachronic tendency We look into the diachronic frequency distribution of each matching head in N-N compounds and their N-PREP-N counterparts. Figure 1 demonstrates how both variants increase in frequency over the period 1665–1869 (especially from the 1750s onwards), with compounds showing a slightly more pronounced trend. Based on this, the diachronic

[4] While the two may appear in free variation in a number of contexts, it is important to note that this may not always be the case. Compare, for example, *tea cup* and *cup of tea* on the one hand to *copper alloy* and *alloy of copper* on the other: the former pair are not semantically equivalent (all such instances are excluded from the analysis), whereas the latter are.

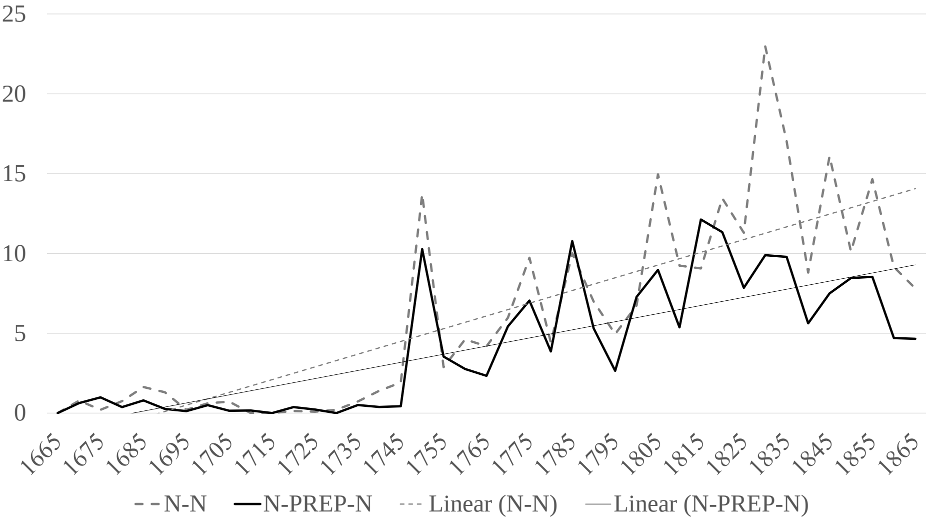

Figure 1: Diachronic development of N-N compounds and their exact N-PREP-N counterparts (frequencies normalized on the basis of nouns in the RSC)

frequency rise expected for compounds in our starting hypothesis appears to be overstated.

Average surprisal We then compare the N-N and N-PREP-N counterparts based on average surprisal, with a word-based model that uses a sliding window of three preceding words for context. For the whole construction, we compare the mean value of average surprisal of each word in the N-N and N-PREP-N counterparts, i.e.

$$\frac{1}{|w|}\sum_{w} AvS(w)$$

where the number of words $|w|$ is 2 for N-N and 3 for N-PREP-N.

This reveals that N-N compounds have an overall higher mean value (6.91 bits) than their N-PREP-N counterparts (4.58 bits). To get a better overview, all 139 variation patterns are also investigated individually. Around 90% of compounds show a higher mean value of average surprisal than their prepositional counterparts, while approximately 10% do not follow this tendency. This seems to vary depending on the head noun involved. In Figure 2, for instance, N-N compounds with the head noun *alloy* have an expectedly higher mean value (bits > 10) than their N-PREP-N counterparts ($6 \leq$ bits ≤ 7). Meanwhile, when looking into compounds with the

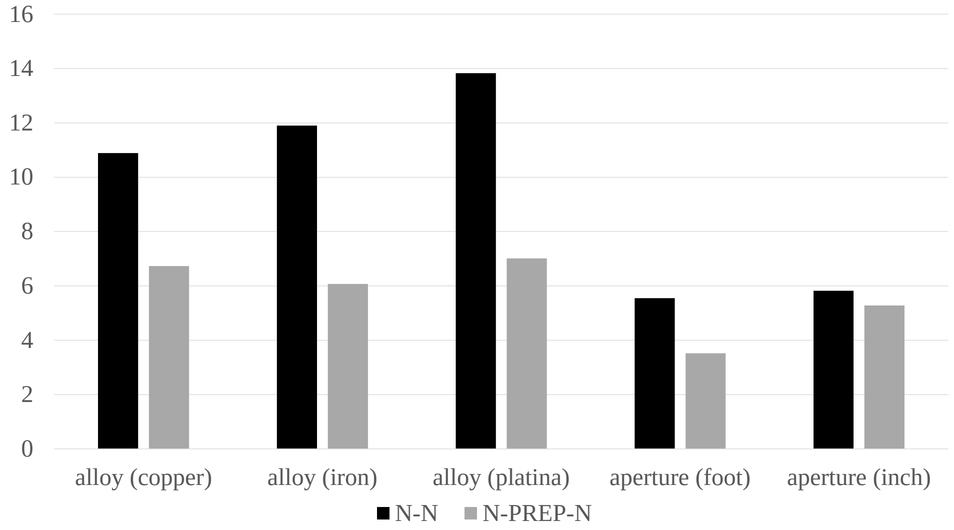

Figure 2: Average surprisal (in bits) for *alloy* and *aperture* counterparts

head noun *aperture* and their exact counterparts, no such tendency can be observed.

Syntagmatic analysis While overall our assumption that N-N compounds have higher encoding density (higher average surprisal) than their N-PREP-N counterparts has been confirmed, we further analyze the syntagmatic context of 10 N-N compounds which – contrary to the overall trend – have a lower mean value of average surprisal than their exact N-PREP-N counterparts. These are *ocean attraction*, *temperature correction*, *inch distance*, *mile distance*, *oxygen gas*, *copper ore*, *copper plate*, *brass ring*, *carbon vapour*, and *copper wire*.

We examine here the effect of (1) lemma representation and (2) surrounding parts of speech on the average surprisal for N-N compounds and their N-PREP-N counterparts.

In terms of lemma representation, first, a plural head noun in N-N compounds (e.g. *attractions* in *ocean attractions*) and N-PREP-N constructions (e.g. *corrections* in *corrections of temperature*) is more likely to carry a higher average surprisal value than its singular counterpart, potentially leading to a higher mean value for the whole construction. Even in cases where a singular head noun (e.g. *gas* in *oxygen gas*, *ore* in *copper ore*) appears to have consistently lower average surprisal than its premodifier, i.e. $p(gas|oxygen, ...) > p(oxygen|...)$, its plural counterpart (e.g. *gases* in *hydrogen and oxygen gases*) reverses the trend by having a higher average surprisal than its directly preceding modifier, i.e. $p(gases|oxygen, ...) < p(oxygen|...)$.

Second, in N-PREP-N constructions involving prepositions other than *of* (e.g. *gas with oxygen*), the preposition (e.g. *with*) tends to be the element with the highest average surprisal. While this may be counterintuitive given the low information content typically associated with function words, it is indicative of how unusual this type of construction is.

With regard to surrounding parts of speech, first, N-N compounds that occur as part of a coordinated structure (e.g. *ocean attraction* in *mountain and ocean attraction*) tend to have a lower mean value than those that do not, as information in the former seems more distributed across the whole structure. By contrast, the N-PREP-N counterpart *attraction of the ocean* – which has a higher mean value – nearly always appears embedded in a prepositional phrase (e.g. *deficiency of attraction of the ocean, ellipticity by the attraction of the ocean*). Second, while an N-N compound such as *steel plate* tends to have a higher mean value when modified by an adjective (e.g. *thin*, *circular*) as opposed to a determiner (e.g. *a*, *the*), the mean value for an N-PREP-N construction such as *plate of copper* typically drops when it is preceded by an adjective rather than a determiner.

If there is a somewhat general tendency to be arrived at, it is that the N-N compound shifts the focus from the head noun to the premodifying noun, whereas the N-PREP-N construction places a higher information value on the head noun. Put simply, the highest average surprisal usually falls on the first noun element in both constructions.

4.2 *Modal Verbs (A2)*

While in the first analysis, we analyzed a field-related feature, in this analysis we consider modal verbs, a feature attributed to the tenor of discourse. According to Leech (2003), the use of modal auxiliaries diminishes over the time period of 1961 to 1992 in English in general (based on the Brown corpus family (Hundt et al., 1999) for written English). There is evidence that this trend started much earlier and also affects scientific writing. Atkinson (1998: 125), for example, notes that in research articles of the Proceedings and Transactions of the Royal Society of London, both prediction and necessity modals decrease over time (prediction modals from 8.6 per 1,000 words in 1675 to 2.3 in 1925; necessity modals from 2.5 in 1825 to 0.9 in 1925). In addition, CLMET (De Smet, 2005), a register-mixed corpus of Late Modern English (1710-1920), exhibits a slight decline of modal verbs from 15,094.73 per million in the period of 1710–1780 to 13,410.27 in the period of 1850–1920. In the RSC, we have observed a similar decline of modal verbs from 2,543.33 to 1,676.17 per million words.

Based on these findings, we inspect the average surprisal of modal verbs exploring contextual factors that might trigger a decline of frequency over time. The unit of modeling for average surprisal is the word with a sliding window of three preceding words:

$$\frac{1}{|w_i|} * -\sum_i log_2 p(w_i|w_{i-1}w_{i-2}w_{i-3})$$

where w_i is the modal verb and w_{i-1} to w_{i-3} are the three preceding words. Thus, we compare the average surprisal values of the modal verb based on different preceding contexts. Note that we look into modal verbs only, excluding semi-modals.

Average surprisal and contextual differences We inspect the average surprisal of modal verbs to see whether there are contextual differences that might lead to a decline in frequency over time. For this, first, we consider the range of average surprisal values of modal verbs in the RSC (see Figure 3). The values range from 0.3 (low in information) to 20 (high in information). The largest quantity of occurrences has average surprisal values between 4 and 9, then there is a drop around 11 to a long tail of relatively rare, high average surprisal values. For closer inspection of the contexts of modal verbs, we take the value of 10 as a cut-off for distinguishing between relatively high (> 10) vs. medium/low average surprisal values (< 10). More specifically, we consider which parts of speech (POS) precede the modal verb in contexts of > 10 vs. < 10. Table 2 shows that for modal verbs with an average surprisal value above 10, common nouns (NN) are the most frequent POS with 31.56% (e.g., *the distance must have been much greater*). For modal verbs with an average surprisal value below 10, instead, personal pronouns (PP) are most frequent with 40.23% (e.g. *and then we shall truly deserve*). This seems to indicate that there is a contextual difference between modal verbs with relatively high (> 10) and low (< 10) average surprisal values. In fact, only 0.86% of modal verbs with a value above 10 are preceded by personal pronouns (e.g. *which I ought to have added*), while 53.61% are nouns (singular and plural common nouns and proper nouns, e.g. *the puncture must be made in the Arms*).

modals > 10			**modals < 10**		
POS	**freq.**	**%**	**POS**	**freq.**	**%**
NN	12,332	31.56	PP	91,385	40.23
,	5,482	14.03	NN	44,604	19.64
NNS	4,613	11.81	,	21,354	9.40
NP	3,912	10.01	NNS	15,788	6.95

NN: singular common noun; NNS: plural common noun; NP: singular proper noun; PP: personal pronoun

Table 2: Parts of speech preceding modal verbs with an average surprisal value above and below 10

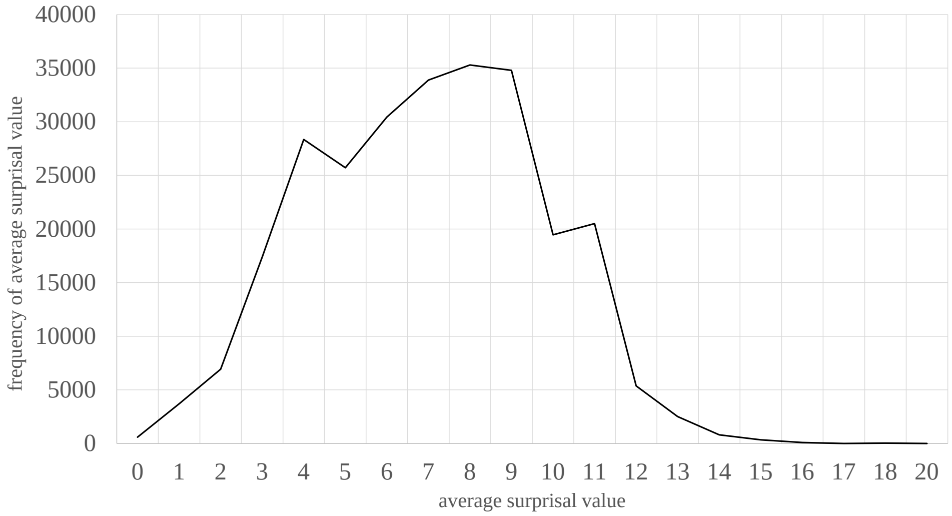

Figure 3: Range of average surprisal for modal verbs across RSC

Considering the range of average surprisal of these two constructions (personal pronoun + modal verb (PP+MV) vs. noun + modal verb (N+MV)), we can see from Figure 4 that the PP+MV has a peak around 4, while the N+MV has a peak around 9. Therefore, a modal verb is more predictable in the context of a preceding personal pronoun than in the context of a preceding noun.

Diachronic tendency based on average surprisal Figure 5 shows the diachronic tendency of the modal verbs in both contexts (PP+MV vs. N+MV). It can be seen that modal verbs used in a more predictive context (PP+MV) are low in information and decrease over time, while modal verbs used in a less predictive context (N+MV) are high in information and increase over time.

4.3 *Part-of-speech Trigrams (A3)*

This analysis focuses on finding possible differences in part-of-speech (POS) trigrams to approximate syntactic patterns that might be involved in diachronic change. For this analysis, we use Kullback-Leibler Divergence (KLD) with the unit of modeling here being the trigram:

$$D_{KL}(T1||T2) = -\sum_i p(trigram_i|T1)log_2\frac{p(trigram_i|T1)}{p(trigram_i|T2)}$$

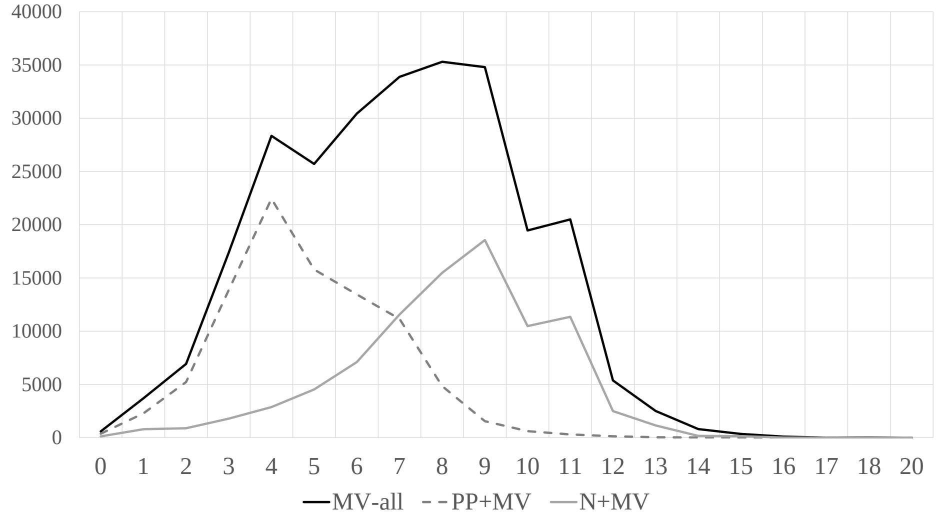

Figure 4: Ranges of average surprisal for all modal verbs vs. personal pronoun + modal verb vs. noun + modal verb across RSC

Comparing the 50-year time periods in the corpus (i.e. *1650*: 1665–1699, *1700*: 1700–1749, *1750*: 1750–1799, *1800*: 1800–1849, *1850*: 1850–1869) to each other, we can observe (1) which POS-trigrams are typical of a particular time period, and (2) which POS-trigrams become more or less typical over time.

We consider only those POS-trigrams that occur at least 20 times in each text. Also, we exclude POS-trigrams consisting of characters constituting sentence markers (e.g. fullstops, colons), brackets, symbols (e.g. equal signs), and words tagged as foreign words. We then create KLD models for each time period against all others (e.g. models for 1700 vs. 1650, 1700 vs. 1750, 1700 vs. 1800 and 1700 vs. 1850).

Typical POS-trigrams of specific time periods To observe POS-trigrams typical of a given time period, we inspect the feature ranking obtained from the KLD values and check whether POS-trigrams of one time period are typical of this time period vs. all other time periods. Thus, for the five time periods, we carry out four comparisons: one time period vs. the other four.

Overall, we observe from Table 3 that the later time periods (1750, 1800 and 1850) have more typical features vs. all others in comparison to the earlier periods (1650, 1700). For the 1650 and the 1700 periods, only one trigram is typical for each (PP.VVZ.DT and PP.VVD.DT, respectively).

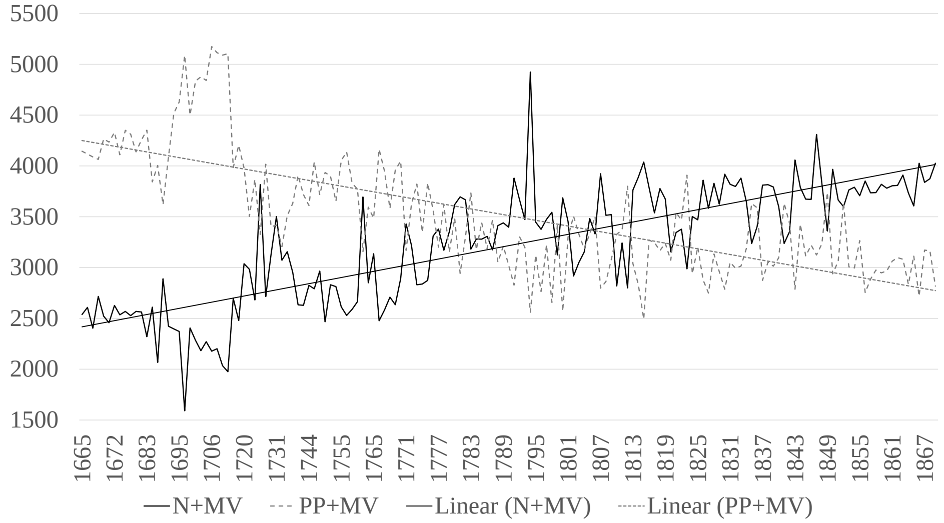

Figure 5: Personal pronoun + modal verb vs. noun + modal verb across RSC

For both periods, the trigram is typical in comparison to later time periods (PP.VVZ.DT is typical of 1650 in comparison to 1750 and 1800; PP.VVD.DT is typical of 1700 in comparison to 1800 and 1850). For the 1750, 1800 and 1850 periods, trigrams are typical in comparison to all other time periods. This seems to indicate that features typical for 1750, 1800 and 1850 are stronger in typicality (as they are typical for a time period towards all others) than typical features of 1650 and 1700 (as they are typical only in comparison to some of the other time periods).

Considering the trigram structure, the typical trigrams for the 1650 and 1700 are quite similar — both are personal pronoun trigrams followed by a verb and a determiner — they just differ in the tense form (VVZ for present vs. VVD for past tense). For the 1750 period, on the other hand, the typical trigrams are quite different from each other: (1) a prepositional phrase (IN.DT.NN), (2) an adverbial phrase (RB.JJR.IN), and (3) a verb phrase with a personal pronoun (PP.RB.VVD). Typical for the 1800, instead, are nominal phrases starting with a determiner and a noun (DT.NN), followed by either a preposition (IN) or the verb *be* in past tense (VBD). These can be taken as indications of the phenomenon of simple relational clauses combined with complex nominal groups as a typical feature of scientific writing pointed out by Halliday (1988). In 1850, again, nominal trigrams are typical, with the top ranking feature being the noun-preposition-noun combination (NN.IN.NN).

POS-trigram	**typical of**	**example**
PP.VVZ.DT	1650	*he gives an/a/the*
PP.VVD.DT	1700	*I found the/a*
IN.DT.NN	1750	*on an inch*
RB.JJR.IN		*little more than*
PP.RB.VVD		*I then took*
DT.NN.IN	1800	*the action of*
DT.NN.VBD		*a paper was*
NN.IN.NN	1850	*inch in diameter*
IN.DT.NN		*of an inch*
DT.JJ.NN		*the same time*

DT: determiner; IN: preposition; JJ: adjective; JJR: comparative adjective; NN: singular common noun; RB: adverb; VBD: verb *be* past tense; VVD: full verb past tense; VVZ: full verb present tense

Table 3: Typical POS-trigrams for the time periods in the RSC

In summary, there seems to be a shift from verbal trigrams with personal pronouns (1650 and 1700) to nominal trigrams (1800 and 1850). 1750 is in between those two groups, as its typical features contain both verbal and nominal trigrams. This is in line with our hypothesis of increasing encoding density as these findings seem to indicate a shift from a verbal to a nominal style, i.e. a shift towards denser encoding.

Diachronic tendencies of POS-trigrams To see whether we can confirm this diachronic tendency, we inspect comparisons between non-adjacent time periods (e.g. 1650 to 1750, 1800, and 1850, excluding the comparison to 1700). We adopt this methodology to show the greatest diachronic differences in POS-trigram use for each period, as we assume that the adjacent time periods would show a greater resemblance with each other.

By inspecting the feature ranking (see Table 4 with examples of the most frequent realizations of the trigrams) for each time period based on the above described selection (i.e. comparison to non-adjacent periods), three observations can be made. First, all POS-trigrams typical of 1650 and 1700 (see rows 1-4 and 5-7) do again include a verbal trigram with a personal pronoun (PP). POS-trigrams typical of 1800 and 1850, instead, are again nominal trigrams (see rows 13-17 and 18-22, respectively). 1750 lies somewhere in between (see rows 8-12), having both nominal and personal pronoun trigrams. Second, diachronically (see again Table 4, right-hand side), we clearly see a decrease in verbal PP-trigrams (typical of 1650, 1700) over time (see rows

1-4 and 5-7) and an increase in nominal trigrams (typical of 1800, 1850; see rows 13-17 and 18-22). Third, if verbal parts of speech are involved, past tense prevails (VVD and VBD (rows 2-4, 5-7, and 11-12) vs. VVZ (row 1)).

In summary, diachronically we can confirm the general tendency towards the use of nominal constructions, while verbal constructions are downplayed over time. This is clearly an indication of a shift from a more situated, involved and personal style (verbal) to a more distant, informational and impersonal style (nominal). This is in line with previous diachronic studies on scientific texts (see e.g. Atkinson (1998), Biber and Finegan (1997), and Moessner (2009) on the Helsinki and ARCHER corpora) which have observed a diachronic tendency towards abstractness and informational production.

5 Conclusion

We have presented an approach to investigate diachronic change in English scientific writing based on information-theoretic models. Concretely, we have proposed to use measures of average surprisal and relative entropy (cf. Section 3). Compared to pure frequency-based accounts, probabilities are calculated *in context* and evaluated with regard to their effects in diachronic change. This provides the following benefits for linguistic analysis:

- conditional probabilities can be used directly for evaluating features (potentially) involved in change,
- explorative analysis for the detection of new features is supported,
- diachronic comparison is facilitated, notably for assessing features in terms of typicality for a given time period relative to others.

We have shown three kinds of analyses using this approach (cf. Section 4), focusing on the hypothesis of increasing encoding density over time (cf. Section 1).

In analysis (A1), we have used average surprisal to assess whether noun-noun compounds carry higher surprisal than their analytic counterparts, the former being more densely encoded compared to the latter. The analysis has confirmed this assumption, as around 90% of the compounds analyzed have exhibited higher average surprisal than their analytic counterparts. In analysis (A2), we have looked at the frequency and usage of modal verbs over time, again using average surprisal. While overall modal verbs decrease in frequency over time, this is dependent on the context of use. We have shown that modal verbs are less surprising (low in information) in the context of a preceding personal pronoun and more surprising (high in information) in the context of a preceding noun.

row no.	**trigram**	**comparison**	**example**	**tendency**	**1650**	**1700**	**1750**	**1800**	**1850**
		1650							
1	PP.VVZ.DT	vs 1750/1800	*he gives an/a/the*	-	562.3	267.3	161.1	174.0	156.2
2	IN.PP.VVD	vs 1800	*as I said*	-	818.7	850.6	581.1	214.6	146.8
3	PP.VVD.PP	vs 1850	*I found it*	-	496.8	618.7	349.3	111.9	48.8
4	PP.VVD.DT		*I found the/a*	-	729.6	1,068.9	822.8	377.5	209.3
		1700							
5	IN.PP.VVD	vs 1800	*as it appeared*	-	818.7	850.6	581.1	214.6	146.8
6	IN.PP.VBD		*that it was*	-	646.9	760.0	528.1	233.2	173.6
7	PP.VVD.DT	vs 1800/1850	*I found the*	-	729.6	1,068.9	822.8	377.5	209.3
		1750							
8	IN.DT.NN	vs 1650/1850	*on an inch*	+	15,426.1	12,159.5	27,941.3	26,089.8	23,549.1
9	NN.IN.NN	vs 1650	*degree of heat*	+	1,336.4	694.0	4,868.1	6,553.5	6,563.6
10	DT.NN.IN		*the quantity of*	+	13,168.1	10,434.5	19,718.2	20,013.3	18,533.0
11	PP.VVD.DT	vs 1850	*I found the*	-	729.6	1,068.9	822.8	377.5	209.3
12	PP.VVD.TO		*it seemed to*	-	350.1	449.1	340.2	177.3	95.0
		1800							
13	NN.IN.NN	vs 1650/1700	*inch in diameter*	+	1,336.4	694.0	4,868.1	6,553.5	6,563.6
14	IN.DT.NN		*of an inch*	+	15,426.1	12,159.5	27,941.3	26,089.8	23,549.1
15	DT.NN.IN		*the action of*	+	13,168.1	10,434.5	19,718.2	20,013.3	18,533.0
16	DT.JJ.NN		*the same time*	+	9,467.8	8,738.1	15,085.4	15,812.9	16,548.1
17	NN.IN.DT		*part of the*	+	12,055.3	10,434.5	17,191.6	18,415.4	18,627.7
		1850							
18	NN.IN.NN	vs 1650/1700/1750	*inch in diameter*	+	1,336.4	694.0	4,868.1	6,553.5	6,563.6
19	IN.DT.NN	vs 1650/1700	*of an inch*	+	15,426.1	12,159.5	27,941.3	26,089.8	23,549.1
20	DT.JJ.NN		*the same time*	+	9,467.8	8,738.1	15,085.4	15,812.9	16,548.1
21	DT.NN.IN	vs 1700	*the number of*	+	13,168.1	10,434.5	19,718.2	20,013.3	18,533.0
22	JJ.NN.IN	vs 1750	*small quantity of*	+	4,839.4	4,420.7	6,990.6	8,272.2	9,241.8

Table 4: Typical POS-trigrams for non-adjacent time periods in the RSC with diachronic tendencies

Only in the first context do modal verbs decrease in frequency over time. This clearly points to a relation between informativeness and linguistic change: less informative usages decreasing in frequency over time and more informative usages surviving or emerging.

The third analysis (A3) has applied relative entropy (Kullback-Leibler Divergence; KLD) to detect new features involved in diachronic change. In contrast to (A1) and (A2), we have used POS-trigrams as a basis for modeling here. This analysis has revealed a shift from verb-based trigrams to noun-based trigrams over time. Again, this supports our hypothesis of increasing encoding density, which is in turn indicative of changes in discourse type (from reporting to expository) and style (from personal/involved to impersonal/informational).

In our ongoing work, we carry out more analyses using other "known" features as well as detect new features involved in change with the methods shown in this paper. Also, we explore other approaches that are promising for diachronic comparison, notably topic models (Fankhauser et al., 2016). Here, we hope to be able to capture discipline-specific language use, e.g. pinpoint academic disciplines in statu nascendi. While the present analyses were focused on linguistic effects of specialization, in future studies we will also look into the possible effects of professionalization/institutionalization, which we assume to lie in increasing linguistic uniformity (cf. Section 1).

Beyond the immediate methodological benefits, adopting the perspective of information in looking at diachronic change may also turn out to be conceptually fruitful: The notion of information promises to allow us to generalize over different, possibly correlated kinds of linguistic changes, it may help us detect phases of diachronic change, and it may itself turn out to be a driving force in language change.

Besides application in the field of diachronic change, the information-theoretic approach described here can be applied to all kinds of comparative studies (comparison of languages, registers etc.) and linguistic features for which contextual differences matter.

Acknowledgments

This research is funded by the German Research Foundation (Deutsche Forschungsgemeinschaft) under grants SFB 1102: Information Density and Linguistic Encoding (www.sfb1102.uni-saarland.de) and EXC-MMCI: Multimodal Computing and Interaction (www.mmci.uni-saarland.de). We are also indebted to Jörg Knappen and Stefan Fischer for their contributions to corpus processing and Peter Fankhauser (IdS Mannheim) for his continuous support in questions of data analysis. Also, we thank the anonymous reviewers for their valuable comments.

References

Aarts, Bas. 1992. *Small Clauses in English: The Nonverbal Types.* Berlin/New York: Mouton de Gruyter.

Asr, Fatemeh Torabi and Vera Demberg. 2013. On the Information Conveyed by Discourse Markers. *Proceedings of the Fourth Annual Workshop on Cognitive Modeling and Computational Linguistics*: 84–93. Sofia, Bulgaria.

Atkinson, Dwight. 1998. *Scientific Discourse in Sociohistorical Context: The Philosophical Transactions of the Royal Society of London, 1675-1975.* New York: Routledge.

Banks, David. 2008. *The Development of Scientific Writing: Linguistic Features and Historical Context.* London: Equinox.

Baron, Alistair and Paul Rayson. 2008. VARD 2: A Tool for Dealing with Spelling Variation in Historical Corpora. *Proceedings of the Postgraduate Conference in Corpus Linguistics*, Aston University, Birmingham, UK, May.

Biber, Douglas. 1988. *Variation across Speech and Writing.* Cambridge: Cambridge University Press.

Biber, Douglas. 2006a *Lexical Bundles in University Teaching and Textbooks*, volume 23 of *Studies in Corpus Linguistics*, chapter 6, 133–175. Amsterdam/Philadelphia: John Benjamins Publishing.

Biber, Douglas. 2006b. *Multi-dimensional Patterns of Variation among University Registers*, volume 23 of *Studies in Corpus Linguistics*, chapter 7, 177–212. Amsterdam/Philadelphia: John Benjamins Publishing.

Biber, Douglas. 2006c. *University Language: A Corpus-based Study of Spoken And Written Registers*, volume 23 of *Studies in Corpus Linguistics.* Amsterdam/Philadelphia: John Benjamins Publishing.

Biber, Douglas and Edward Finegan. 1997. Diachronic Relations among Speech-based and Written Registers in English. In Terttu Nevalainen and Leena Kahlas-Tarkka (eds.), *To Explain the Present: Studies in the Changing English Language in Honour of Matti Rissanen*, 253–276. Helsinki: Société Néophilologique.

Biber, Douglas and Bethany Gray. 2016. *Grammatical Complexity in Academic English: Linguistic Change in Writing.* Cambridge: Cambridge University Press.

Biber, Douglas and Bethany Gray. 2013. Nominalizing the Verb Phrase in Academic Science Writing. In Bas Aarts, Joanne Close, Geoffrey Leech, and Sean Wallis (eds.), *The Verb Phrase in English*, 99–132. Cambridge: Cambridge University Press.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. Harlow: Longman.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3: 993–1022.

Cockburn, Alistair. 2001. *Agile Software Development*. Boston: Addison-Wesley Professional.

Crocker, Matthew W., Vera Demberg, and Elke Teich. 2015. Information Density and Linguistic Encoding (IDeaL). *KI - Künstliche Intelligenz*, 30: 77–81.

CWB. 2016. The IMS Open Corpus Workbench. Downloadable at http://www.cwb.sourceforge.net.

De Smet, Hendrik. 2005. A Corpus of Late Modern English. *ICAME Journal* 29: 69–82.

Evert, Stefan and Andrew Hardie. 2011. Twenty-first Century Corpus Workbench: Updating a Query Architecture for the New Millennium. *Proceedings of the Corpus Linguistics Conference*. Birmingham, UK. Downloadable at http://eprints.lancs.ac.uk/62721/.

Fankhauser, Peter, Jörg Knappen, and Elke Teich. 2014. Exploring and Visualizing Variation in Language Resources. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*: 4125–4128. Reykjavik, Iceland, May.

Fankhauser, Peter, Jörg Knappen, and Elke Teich. 2016. Topical Diversification over Time in the Royal Society Corpus. *Proceedings of Digital Humanaties (DH)*, Krakow, Poland, July.

Firth, John Rupert. 1957. *Papers in Linguistics 1934-1951*. London: Oxford University Press.

Genzel, Dmitriy and Eugene Charniak. 2002. Entropy Rate Constancy in Text. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*: 199–206. Stroudsburg, PA, USA. Downloadable at http://dx.doi.org/10.3115/1073083.1073117.

Halliday, M.A.K. 1985. *Written and Spoken Language.* Melbourne: Deakin University Press.

Halliday, M.A.K. 1988. On the Language of Physical Science. In Mohsen Ghadessy (editor), *Registers of Written English: Situational Factors and Linguistic Features*, 162–177. London: Pinter.

Halliday, M.A.K. and Ruqaiya Hasan. 1985. *Language, Context, and Text: Aspects of Language in a Social-semiotic Perspective.* Oxford: Oxford University Press.

Halliday, M.A.K. and J.R. Martin. 1993. *Writing Science: Literacy and Discursive Power.* London: Falmer Press.

Hardie, Andrew. 2012. CQPweb – Combining Power, Flexibility and Usability in a Corpus Analysis Tool. *International Journal of Corpus Linguistics*, 17(3): 380–409.

Hundt, Marianne, Andrea Sand, and Rainer Siemund. 1999. *Manual of Information to Accompany The Freiburg LOB Corpus of British English (FLOB).* Freiburg: Department of English, Albert-Ludwigs-Universität Freiburg.

Hundt, Marianne, David Denison, and Gerold Schneider. 2012. Relative Complexity in Scientific Discourse. *English Language and Linguistics* 16: 209–240.

Jaeger, T. Florian. 2011. Corpus-based Research on Language Production: Information Density and Reducible Subject Relatives. In Emily M. Bender and Jennifer E. Arnold (eds.), *Language From a Cognitive Perspective: Grammar, Usage, and Processing*, 161–197. Standford: CSLI Publishers.

Kermes, Hannah, Stefania Degaetano-Ortlieb, Ashraf Khamis, Jörg Knappen, and Elke Teich. 2016. The Royal Society Corpus: From Uncharted Data to Corpus. *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC).* Portoroz, Slovenia, May.

Leech, Geoffrey. 2003. Modality on the Move: The English Modal Auxiliaries 1961-1992. In Roberta Facchinetti, Manfred Krug, and Frank Palmer (eds.), *Modality in Contemporary English. Topics in English Linguistics* 44, 223–240. Berlin: Mouton de Gruyter.

Leech, Geoffrey and Nicholas Smith. 2009. *Corpus Linguistics: Refinements and Reassessments*, chapter Change and Constancy in Linguistic Change: How Grammatical Usage in Written English Evolved in the Period 1931-1991, 173–200. Amsterdam/New York: Rodopi.

Manning, Christopher D. and Hinrich Schütze. 2001. *Foundations of Statistical Natural Language Processing.* Cambridge/London: The MIT Press.

Moessner, Lilo. 2009. The Influence of the Royal Society on 17th-century Scientific Writing. *ICAME Journal* 33: 65–88.

Sayeed, Asad, Stefan Fischer, and Vera Demberg. 2015. Vector-space Calculation of Semantic Surprisal for Predicting Word Pronunciation Duration. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*: 763–773. Beijing, China, July. Downloadable at URL http://www.aclweb.org/anthology/P15-1074.

Schmid, Helmut. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of the International Conference on New Methods in Language Processing*: 44–49. Manchester, UK.

Schmid, Helmut. 1995. Improvements in Part-of-Speech Tagging with an Application to German. *Proceedings of the ACL SIGDAT-Workshop.* Kyoto, Japan.

Shannon, Claude E. 1949. *The Mathematical Theory of Communication.* Urbana/Chicago: University of Illinois Press, 1983 edition.

Taavitsainen, Irma and Päivi Pahta. 2010. *Early Modern English Medical Writing: Corpus Description and Studies.* Amsterdam/Philadelphia: John Benjamins.

Taavitsainen, Irma, Peter M. Jones, Päivi Pahta, Turo Hiltunen, Ville Marttila, Maura Ratia, Carla Suhr, and Jukka Tyrkkö. Medical Texts in 1500-1700 and the Corpus of Early Modern English Medical Texts. In Irma Taavitsainen and Päivi Pahta (eds.), *Medical Writing in Early Modern English*, 9–29. Cambridge: Cambridge University Press.